Q-learning Assisted Energy-Aware Traffic Offloading and Cell Switching in Heterogeneous Networks

Attai Ibrahim Abubakar, Metin Ozturk, Sajjad Hussain and Muhammad Ali Imran
School of Engineering, Unversity of Glasgow, Glasgow, UK
{a.abubakar.1, m.ozturk.1}@research.gla.ac.uk, {Sajjad.Hussain, Muhammad.Imran}@glasgow.ac.uk

***Abstract*—Cell switching has been identified as a major approach to significantly reduce the energy consumption of Heterogeneous Networks (HetNets). The main idea behind cell switching is to turn off idle or lightly loaded Base Stations (BSs) and to offload their traffic to neighbouring active cell(s). However, the impact of the offloaded traffic on the power consumption of the neighbouring cell(s) has not been studied sufficiently in the literature, thereby leading to the development of sub-optimal cell switching mechanisms. In this work, we first considered a Control/Data Separated Architecture (CDSA) with a macro cell serving as the Control Base Station (CBS) and multiple small cells as Data Base Stations (DBS). Then, a Q-learning assisted cell switching algorithm is developed in order to determine the small cells to switch off by considering the increase in power consumption of the macro cell due to offloaded traffic from the sleeping cells. The capacity of the macro cell is also taken into consideration to ensure that the Quality of Service (QoS) requirements of users are maintained. Simulation results show that the proposed cell switching algorithm can achieve up to 50% reduction in the total energy consumption of the considered HetNet scenario.**

I. Introduction

Multi-tier Heterogeneous Network (HetNet) where massive number of small cells are deployed under the coverage of macro cell(s) have been identified as one of the major approaches to enhance the capacity of 5G networks [1]. Small cells have the advantage of proximity to user resulting in reduced propagation loss as well as improved signal-to-interference-plus-noise ratio (SINR) and lesser delay. However, massive deployment of small cells leads to increase in the energy consumption of wireless networks, which in turn increases CO_2 emission and network operating costs [2].

Base Stations (BSs) have been identified as the major contributor to the total energy consumption in cellular networks [3], hence minimizing the energy consumption of BSs can significantly reduce the overall energy consumption of the network. Since 5G promises a thousand-fold increase in data rate, a thousand-fold increase in energy efficiency needs to be achieved in order to maintain energy consumption at its current level [2].

The temporal and spatial variation of traffic load in cellular networks due to varying user distribution and mobility patterns results in most BSs been either idle (i.e. serving no user) or under-utilized (i.e. serving very few users) for most part of the day which amounts to energy wastage. Hence, a significant amount of energy saving can be obtained by dynamically turning off idle and underutilized cells during periods of low traffic load.

In conventional HetNets, switching off small cells is quite challenging as it often results in the existence of coverage holes and hence degradation in the QoS of users originally associated with the deactivated cell(s). This is not the case in the HetNet with Control/Data Separated Architecture (CDSA) where the macro BS provides continuous coverage, signalling functions and low data rate services while the small cells handles high data rate transmissions. With coverage always guaranteed by the macro BS, the small cells can be easily turned off/on without affecting the QoS of users [4].

In addition to energy savings, the QoS requirements of users must be also be guaranteed using techniques such as traffic offloading and user association. Traffic offloading for energy efficient network operation entails transferring the traffic load of sleeping BS(s) to other active BS(s). Three categories of traffic offloading have been identified in literature. First, vertical offloading [5], [6] where the traffic load of the sleeping cell(s) is transferred to the macro cell, secondly, horizontal offloading [7], [8] which involves offloading the traffic to the neighbouring cell(s) and lastly, joint traffic offloading [9], [10] where both vertical and horizontal traffic offloading techniques are employed. The use of reinforcement learning for traffic offloading and cell switching has also been considered in [7], [8].

A Transfer Actor Critic (TACT) model was applied in [8] while in [7] a centralized and decentralized Q-learning algorithm was used to determine the cell switching strategy. However, none of these works considered the increase in transmission power of the macro BS due to offloaded traffic from the sleeping small cell when developing their cell switching mechanisms thereby leading to sup-optimal cell switching designs. Moreover, reinforcement learning have mostly been applied for horizontal traffic offloading and cell switching in conventional HetNets but to the best of our knowledge, it has not yet been applied for vertical traffic offloading in CDSA. Therefore, we propose a reinforcement learning based cell switching scheme that employs vertical traffic offloading to turn off lightly loaded small cell(s) in a HetNet with CDSA during periods of low traffic load. We consider the additional energy consumption of the macro BS due to offloaded traffic as well as the capacity of the macro BS in order to develop an efficient switching mechanism.

The rest of the paper is organized as follows. Section II

discusses related works, while the system model is introduced in Section III. Section IV presents the optimization objective and constraints. The Q-learning framework for small cell switching is presented in Section V, while Section VI presents the performance evaluation of the proposed learning algorithm. Section VII concludes the paper.

II. Related works

The design of efficient cell sleeping techniques with traffic offloading has attracted much research attention lately. A novel small cell wake up scheme was developed in [5], where the small cells offload their traffic to macro cell before entering into sleep mode while an optimal number of small cells are woken up to accommodate the increase in traffic load during peak traffic periods. In [6], an analytical model to determine the number of small cells that can be switched off with vertical traffic offloading was proposed using two sleeping schemes (random and repulsive scheme). In the random scheme, the small cells have equal probability of being turned off while in the repulsive scheme, only the small cell(s) closest to the macro BS are turned off. The use of reinforcement learning framework to optimise cell switching strategy was considered in [7], [8], [11], [12]. The author in [11], [12] apply Q-learning algorithm for adaptive sleep mode management in order to optimize energy consumption of BSs in 5G homogeneous network deployment. In [7], centralized and decentralized Q-learning algorithm was developed to optimize the traffic offloading and small cell switch off process. A Transfer Actor-Critic (TACT) model to optimize the dynamic switching off/on of small cells in order to match traffic load with energy consumption in a HetNet was developed in [8].

Although previous works [7], [8] have applied reinforcement learning for cell switching with horizontal traffic offloading, none have applied it to vertical traffic offloading. In addition, HetNet scenario with CDSA has never been considered in reinforcement learning aided traffic offloading techniques. Also, in designing vertical traffic offloading and small cell switching schemes, one of the factors that needs to be considered is the increase in transmission power of the macro cell in case of traffic offloading from small cells. In other words, energy saving obtained from switching off the lightly loaded small cells should be compared with the increase in the transmit power of the macro cell in order to make the cell switching process valid and more reasonable. This criteria is often overlooked in many studies [5], [6] when designing small cell switching schemes in both the conventional HetNet and HetNet with CDSA.

Therefore, we develop a Q-learning algorithm for vertical traffic offloading and small cell switching to minimize energy consumption in HetNet with CDSA while taking into cognizance the capacity of the macro cell as well as the incremental power consumption on the macro cell due to traffic offloading from sleeping small cells.

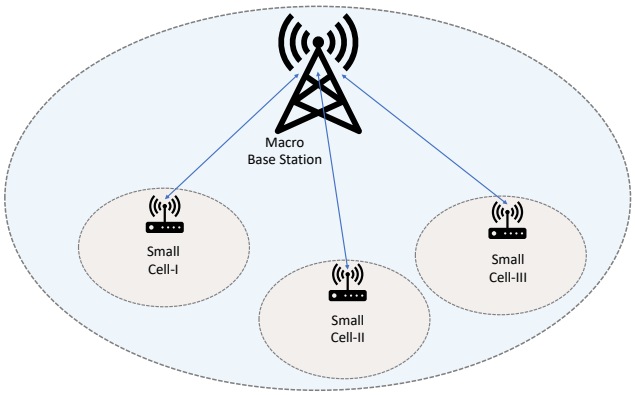


Fig. 1. Network model comprising a macro BS and three small cells in HetNet deployment with CDSA

III. System Model

A. Network Model

We consider a two-tier HetNet consisting of a macro cell and three small cells as shown in Fig. 1, with separated control plane and data plane where the macro cells and small cells operate in dedicated frequency bands. The macro cell is responsible for providing coverage, signalling as well as low data rate services while the small cells provide high capacity in hotspot locations and are linked to the macro cell using the backhaul. Small cells are switched off during low traffic load periods and their traffic offloaded to their associated macro cell provided there is enough capacity in the macro cell to accommodate the offloaded traffic load.

B. Power Consumption Model of HetNet

The Earth model [3] for determining the total power consumption of base station is adopted and is expressed as:

$$P_{sum} = \begin{cases} P_{\mathrm{c}} + \delta_y P_{\mathrm{tx}} & 0 < P_{\mathrm{tx}} < P_{\max} \\ P_{\mathrm{s}} & P_{\mathrm{tx}} = 0, \end{cases} \tag{1}$$

where P_{sum} is the total power consumption of a BS, δ_{y} is the slope of the load dependent power consumption, P_{c} denotes the constant power consumption component of the BS when in operation, P_{s} is the power consumed by the BS when in sleep mode, $P_{\mathrm{tx}}, P_{\max}$ are the instantaneous and maximum transmit power of the BS respectively.

We assume a network consisting of BSs (both macro and small cells) each having limited number of resource blocks (RB) and that both macro cell and small cell have the same number of RBs. We model the load profile of each BS as the proportion of RBs occupied per minute over a 24 hour period. Hence, for a BS having (N_{T}) total number of RBs with (N_{u}) number of RBs occupied per minute, the load (ρ_i) of the BS per minute as well as the relation between the load (ρ_i), the instantaneous (P_{tx}) and maximum transmit power $(P_{\max})$ can be expressed as [3]:

$$\rho_i = \frac{N_{\mathrm{u}}}{N_{\mathrm{T}}} = \frac{P_{\mathrm{tx}}}{P_{\max}}. \tag{2}$$

$$P_{\text{tx}} = \rho_i . P_{\text{max}}, \tag{3}$$

where $i = \{1, 2, 3, ..., \text{n}\}$, and i is in minutes

Combining (1) and (3) and assuming that P_{s} is zero when the BS is in sleep mode, the total power consumption of a macro BS can be expressed as:

$$P_{\text{m}} = P_{\text{c}}^{\text{m}} + \delta_{\text{m}} \rho_i^{\text{m}} P_{\text{max}}^{\text{m}}, \tag{4}$$

where P_{m} denotes the total power consumption of a macro BS, P_{c}^{m} denotes the constant power consumption component of the BS when the BS is in operation, δ_{m} is the load dependent component of power consumption of the macro BS, ρ_i^{m} is the load of the macro BS per minute.

The total power consumption of a small cell is given as:

$$P_{\text{s}}^{k} = P_{\text{c}}^{\text{s}} + \delta_{\text{s}} \rho_i^{s} P_{\text{max}}^{\text{s}}, \tag{5}$$

where P_{s}^{k} denotes the total power consumption of a small BSs, $k = \{1, 2, 3, ...\text{K}\}$, is the number of small cells, P_{c}^{s} denotes the constant power consumption component of small BS in active mode, δ_{s} is the load dependent component of power consumption of the small BS, ρ_i^{s} is the load of the a small cell at every minute.

The total power consumption of the HetNet is the sum of the power consumption of the macro cell(s) and all the smalls cells under its coverage. It can be written as:

$$P_{\text{HetNet}} = P_{\text{m}} + \sum_{k=1}^{K} P_{\text{s}}^{k}, \tag{6}$$

where P_{HetNet}, P_{m} and P_{s}^{k} is the total power consumption of the HetNet, the power consumption of the macro BS as well as that of the k-th small BS respectively.

IV. Problem Formulation

The aim is to determine the optimal strategy to switch off lightly loaded small cells during low traffic periods that will minimise the total power consumption of the HetNet while considering the QoS requirement of offloaded traffic which is availability of capacity. Therefore, our optimization objective can be defined as:

$$\begin{aligned} \min_{\psi \in \Psi} \quad & \text{P}\big(\psi\big) \\ \text{s.t.} \quad & \sum N_{\text{s}}^{\text{off}} < \big(N_{\text{m}}^{\text{T}} - N_{\text{m}}^{\text{U}}\big) \\ & \sum P_{\text{s-off}}^{k} > \Delta P_{\text{m}}, \end{aligned} \tag{7}$$

where Ψ is the set of all possible small cell switching strategies. $\text{P}\big(\psi\big)$ is the expected power consumption of the HetNet using any switching strategy ψ. The first constraint is the capacity constraint, which implies that the number of RBs required to offload the traffic of sleeping small BS(s), $\sum N_{\text{s}}^{\text{off}}$ must be less than the available number of RBs in the macro BS, where N_{m}^{T} and N_{m}^{U} are the total number of resource blocks and utilized resource blocks in the macro cell respectively. The second constraint is the dynamic power (power consumption due to transmission) constraint which implies that the power consumption gain $\sum P_{\text{s-off}}^{k}$ obtained by switching off small BS(s) must be greater than the increase in power consumption in the macro cell, ΔP_{m} as a result of additional load from sleeping small BS(s). We develop a reinforcement learning based small cell switching and traffic offloading mechanism in the next session to optimize energy consumption in the HetNet.

V. Proposed Methodology

We propose a reinforcement learning algorithm, a set of machine learning algorithms, to implement the small cell switching operation. Reinforcement learning is a risk and reward kind of learning whereby the agent (or macro cell) gets information from the environment and then tries to take action and is rewarded or penalized depending on whether the action taken is right or wrong. Reinforcement learning is applied in this work due to its suitability to handle this kind of tasks that involve making decisions out of a wide-range of options [13]. As a illustration in our study, the macro cell interacts with the network environment, obtains information about the traffic loads levels of the small cells through its backhaul connection with them and then decide which combination of small cells to switch off per time. Hence, reinforcement learning is able to cope with the requirement for solving this kind of problem because it can adapt to changing environment through learning and then decide the action that would yield the best desired performance.

In this work, we adopt Q-learning algorithm [14]. Q-learning is one of the most popular reinforcement algorithms, and has a proven capability of working in dynamic environments [15]. There are six main components in Q-learning: (i) agent, (ii) environment, (iii) action, (iv) state, (v) reward/penalty, and (vi) action-value table. Agent takes actions by interacting with a given environment in order to maximize the reward or minimize the penalty. After each action that the agent takes, resulting state and reward/penalty are evaluated. Then, the action-value table, which stores the rewards/penalties for all the possible actions and states, are updated according to following rule:

$$\begin{aligned} Q(s_t, a_t) := Q(s_t, a_t) + \\ \lambda\big[\gamma_{t+1} + \phi \min_{a}(Q(s_{t+1}, a)) - Q(s_t, a_t)\big], \end{aligned} \tag{8}$$

where s_t and s_{t+1} are the current and next states, respectively. γ_{t+1} is the expected penalty for the next step and a_t is the taken action, where a is the set of all possible actions. λ is a learning rate while ϕ is a discount factor. min function in (8) should be converted to max function to make the update policy suitable for the reward-based framework, which includes a reward function, Γ, instead of γ.

Q-learning is an off-policy method, meaning that it follows different policies in determining the next action and updating the action-value table. Although ϵ-greedy is the base policy, π policy, where $\epsilon > 0$, is followed in selecting the next action, while μ policy, where $\epsilon = 0$, is followed in updating the

action-value table. Moreover, Q-learning is a model-free approach, where the agent does not have a prior knowledge about the environment; instead it interacts with the environment by taking actions.

The motivation for using Q-learning in this work stems from the fact that as a model free learning algorithm [14], it is suitable for application in dynamic environments whose statistics continually changes such as the traffic loads of BSs in a HetNet and it has low computational overhead compared to other cell switching heuristic algorithms which mainly employ exhaustive search techniques. Hence, it can lead to a more robust and scalable implementation of BS switching even when the network size is large. It has also been proven to converge to optimal solution most of the time [16]. In this work, we assume a simple HetNet model with 1 macrocell and 3 small cells such that the state space will be small enough to apply a simple look up table (Q table), which is updated for every state action pair.

In designing the small cell switching mechanism, our goal is to find the best switching strategy i.e., select the best set of small cell(s) to switch off out of all possible set of small cells. This is known as the optimal policy in reinforcement learning. We consider a simple HetNet deployment scenario as a representative case comprising 1 macro BS and 3 small BSs which can later be generalized with more BSs. The environment is the traffic loads levels of the small cells. The state is related to our optimization constraint which is availability of capacity in the macro cell for traffic offloading. Two states, δ_1 and δ_2, are then described as follows:

$$\begin{cases} \delta_1, & C_m < C_{ol}, \\ \delta_2, & C_m \geqslant C_{ol}, \end{cases} \tag{9}$$

where the first state is when the capacity constraint is not satisfied and the second state is when the capacity constraint is satisfied. The penalty function, γ, is designed to be the total power consumption of the network as in (6):

$$\gamma(a) = P_{\text{HetNet}}(P_{\text{m}}, P_{\text{s}}^{\text{k}}). \tag{10}$$

There are eight possible action sets that the agent can take. These actions correspond to each policy, that is, the set of small BS(s) that can be switched off at a given time instant.

The proposed Q-learning algorithm is provided in Algorithm 1, where w is the window size for resetting the action-value table.

Algorithm 1: Proposed Q-Learning Algorithm

Input : Traffic loads of macro and small cells
Output: Small cells to be switched off

Initialize $Q(s, a) := 0$;
for *every episode* **do**
  if $episode \equiv 0$ *(mod* w*)* **then**
    Initialize $Q(s, a) := 0$;
  end
  for *iterations* **do**
    Determine the current state using (9);
    Take an action;
    Calculate penalties through (10);
    Go to the next state;
    Update the action-value table with (8).
  end
end

VI. Performance Evaluation

In this section, we evaluate the performance of the proposed Q-learning based cell switching algorithm. The network parameters used for the simulation were obtained from [3] and are listed in Table I. The learning rate, λ, was set to 0.3 while the discount factor, ϕ, was set to 0.9 [15]. The simulation environment comprises a macro BS and three smalls cells. We consider a scenario where horizontal offloading among the small cells is not possible as their footprints do not overlap. As such, vertical offloading from small cells to macro cell is considered in this work. Also the macro cell controls the switching off/on operation of the small cells under its coverage. The network is monitored over a 24 hour period with 1 minute resolution, meaning that the switching operation is performed every minute. We compare the energy consumption of the HetNet with and without Q-learning at different BSs traffic load. With no Q-learning, no switching mechanism is implemented, hence all the BSs are active irrespective of their traffic load level. Then, the energy consumption gain achieved by implementing the proposed Q-learning based cell switching algorithm is quantified.

TABLE I
SIMULATION PARAMETERS

Parameter	Value
Bandwidth	20 MHz
Number of RBs per macro cell	100
Number of RBs per small cell	100
$P_{\text{max}}^{\text{m}}$, $P_{\text{max}}^{\text{s}}$	20 W, 6.3 W
P_{c}^{m}, P_{c}^{s}	130 W, 56 W
δ_{m}, δ_{s}	4.7, 2.6
Number of iterations	100

The load of the small cells, ρ_{s}, are generated using a uniform random distribution and can be depicted as $\rho_{\text{s}} \in [1, m_{\text{s}}]$, where m_{s} is the normalized maximum load level of the small cell. Similarly, a uniformly distributed random traffic load of the macro BS, ρ_{m}, is also generated using $\rho_{\text{m}} \in [1, m_{\text{m}}]$, where m_{m} is the normalized maximum load level of the macro cell. For each simulation, we specify the maximum load level of the macro BS while the load of the small cells are continually varied. The energy consumption of the HetNet with and without learning is depicted in Fig. 2. It can be be observed that there is a significant reduction in the total power consumption of the HetNet with the application of the developed Q-learning-based cell switching algorithm. This is because Q-learning is able select the optimal set of small

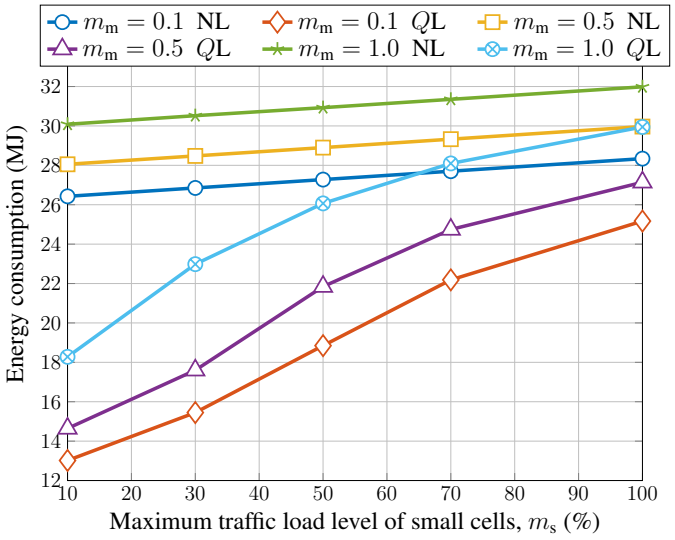


Fig. 2. Total HetNet energy consumption with and without Q-learning where NL represents energy consumption without Q-learning while QL represents energy consumption with Q-learning and m_m is the normalized maximum load of the macro BS. $\alpha = 0.1$ and $w = 50$.

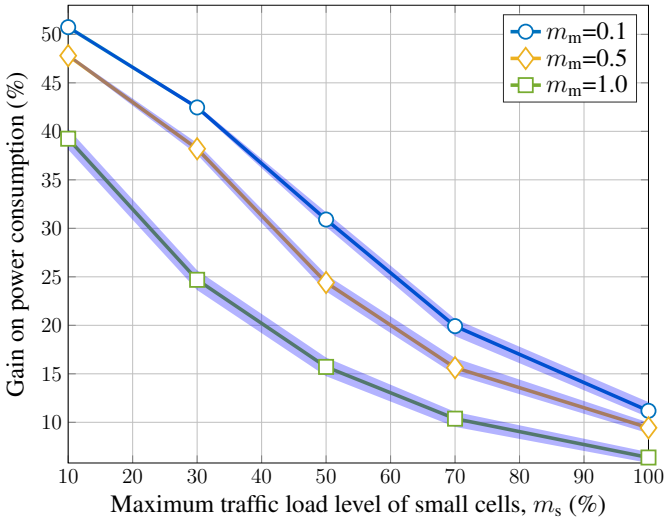


Fig. 3. Energy consumption gain with Q-learning (i.e. percentage reduction in HetNet energy consumption with Q-learning). $\alpha = 0.1$ and $w = 50$. Note that while the shaded areas in the figure show the confidence levels (minimums and maximums of 100 runs) of the findings, the straight lines with markers represent the averages of the runs.

cell(s) to be switched off per time thereby enabling the HetNet to operate with minimal energy consumption.

Fig. 2 also shows that the energy consumption of the HetNet increases as the traffic load of the small cells increases. With increasing traffic load on the small cell, the opportunity for offloading traffic to the macro BS reduces due to availability of limited resources, therefore more small cells have to be left in active mode in order to sustain the increased network traffic load. As a result, the HetNet has to operate at higher energy consumption when the traffic load increases. Also from Fig. 2, the lower the maximum load of the macro BS, (m_m), the lesser the energy consumption since there will be provision to switch off more small cells but higher m_m values results in higher energy consumption.

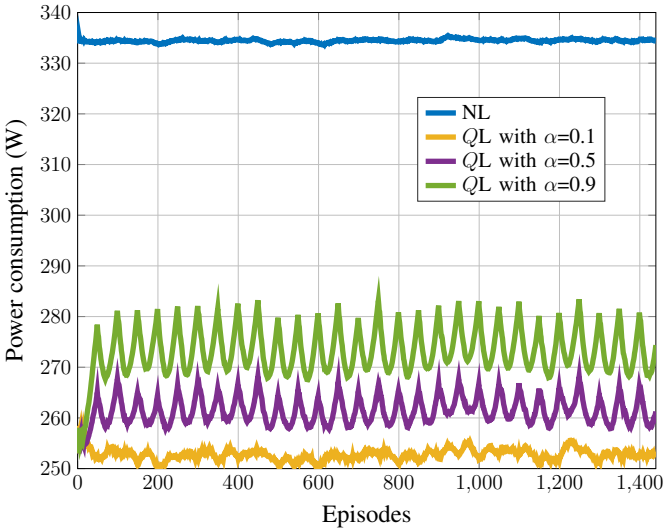
Fig. 3 presents the gain in energy consumption as well as the confidence levels of the results obtained when Q-learning is implemented. Since the findings in Fig. 3 are obtained using the results of Fig. 2, the confidence levels are only presented in Fig. 3 for the sake of simplicity of the presentation. The energy consumption gain is the percentage reduction in energy consumption of the HetNet due to the application of the proposed framework. Similar to the observation in Fig. 2, the power consumption gain reduces with increasing small cell load, as the probability of switching off smalls cell reduces with increasing traffic load.

Fig. 3 also shows that higher energy consumption gain is obtained with lesser m_m values but the gain decreases as the value of m_m increases because more switching opportunities exits when the maximum load of the macro BS is low. The simulation results reveals that an energy consumption gain of up to 50% can be achieved with the proposed Q-learning framework.

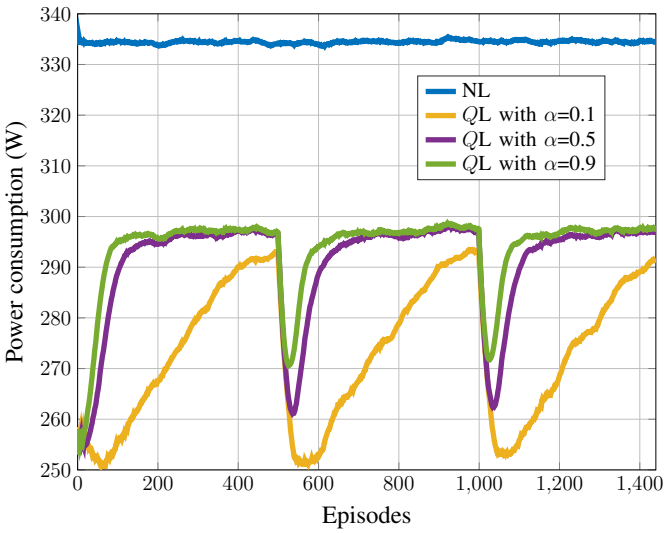
The experimental proof of convergence of the proposed Q-learning algorithm and the impact of learning rate, α, and the window size, w, for initializing the action-value table are shown in Fig. 4. The main idea of initializing the action-value table with w is to decrease the computational expense of Q-learning implementation. Ideally, the action-value table should be initialized once at the beginning of the implementation, and kept the same until the end in order to reduce the computational cost. When building the action-value table, environment learning takes some time in the beginning; however, once the environment is learnt, minor changes in the built action-value table would be enough for Q-learning to adapt itself to new conditions. Nonetheless, this is only the case for gradually changing environments, where Q-learning adapts itself easily. Since we determine the traffic loads of the macro cell and small cells in a random manner, it results in abrupt changes in the environment of interest, making the built action-value table no more valid, as the experienced environment might be significantly different from the learnt one.

Initializing the action-value table at every episode could be an approach for this kind of abruptly changing environments; however, it comes with the expense of computational burden, since it makes Q-learning try to learn the environment at each episode. Thus, instead of initializing the action value-table at each episode, it could be initialized at every w episodes in order to save from the computational cost. However, there is a trade-off between the performance of Q-learning and the computational cost, making the selection of proper α and w quite critical.

There are two main takeaways that can be inferred from Fig. 4: 1) Comparing Fig. 4a and Fig. 4b, smaller w values give better results in terms of the performance of Q-learning, as the action-value table learnt just after the initialization is not valid for upcoming episodes due to abrupt changes in the environment, resulting in performance degradation. 2) Fig. 4a



(a) Window size of $w = 50$



(b) Window size of $w = 500$

Fig. 4. Performance impacts of learning rate, α and the window size for resetting the action-value table, w, on Q-learning convergence. The maximum traffic loads for both macro cell and small cells (m_{m} and m_{s}) are set to 0.5. The results are the averages of 100 runs.

reveals that having smaller α value is better given that Q-learning starts focusing on the new observations more with decreasing α. Therefore, in this work, w and α are selected as 50 and 0.1, respectively.

VII. Conclusion

In this paper, we developed a Q-learning algorithm to minimize the total power consumption of a two-tier HetNet with CDSA. The Q-learning based cell switching algorithm is able to select the optimal set of small cells to be switched off in order to maximize energy saving in the HetNet. The increase in power consumption in the macro cell due to offloaded traffic from sleeping smalls cells was also considered when developing the switching mechanism. The result of the simulation reveals an energy savings gain of about 50% while ensuring that the QoS of users is maintained. A simple network deployment was considered in this study, future work would consider a more complex network scenario and make use of more realistic traffic data in order to develop a cell switching algorithm that is applicable to practical network scenarios.

Acknowledgement

This work was supported by EPSRC Global Challenges Research Fund the DARE Project under Grant EP/P028764/1.

References

[1] A. Damnjanovic, J. Montojo, Y. Wei, T. Ji, T. Luo, M. Vajapeyam, T. Yoo, O. Song, and D. Malladi, "A survey on 3GPP heterogeneous networks," *IEEE Wireless communications*, vol. 18, no. 3, pp. 10–21, 2011.

[2] S. Buzzi, C.-L. I, T. E. Klein, H. V. Poor, C. Yang, and A. Zappone, "A survey of energy-efficient techniques for 5G networks and challenges ahead," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 4, pp. 697–709, apr 2016.

[3] G. Auer, V. Giannini, C. Desset, I. Godor, P. Skillermark, M. Olsson, M. Imran, D. Sabella, M. Gonzalez, O. Blume, and A. Fehske, "How much energy is needed to run a wireless network?" *IEEE Wireless Communications*, vol. 18, no. 5, pp. 40–49, oct 2011.

[4] A. Mohamed, O. Onireti, M. A. Imran, A. Imran, and R. Tafazolli, "Control-data separation architecture for cellular radio access networks: A survey and outlook," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 1, pp. 446–465, 2016.

[5] Y.-B. Lin, L.-C. Wang, and P. Lin, "SES: A novel yet simple energy saving scheme for small cells," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 9, pp. 8347–8356, sep 2017.

[6] S. Zhang, J. Gong, S. Zhou, and Z. Niu, "How many small cells can be turned off via vertical offloading under a separation architecture?" *IEEE Transactions on Wireless Communications*, vol. 14, no. 10, pp. 5440–5453, oct 2015.

[7] X. Chen, J. Wu, Y. Cai, H. Zhang, and T. Chen, "Energy-efficiency oriented traffic offloading in wireless networks: A brief survey and a learning approach for heterogeneous cellular networks," *IEEE Journal on Selected Areas in Communications*, vol. 33, no. 4, pp. 627–640, apr 2015.

[8] R. Li, Z. Zhao, X. Chen, J. Palicot, and H. Zhang, "TACT: A transfer actor-critic learning framework for energy saving in cellular radio access networks," *IEEE Transactions on Wireless Communications*, vol. 13, no. 4, pp. 2000–2011, apr 2014.

[9] R. Tao, W. Liu, X. Chu, and J. Zhang, "An energy saving small cell sleeping mechanism with cell range expansion in heterogeneous networks," *IEEE Transactions on Wireless Communications*, pp. 1–1, feb 2019.

[10] X. Xu, C. Yuan, W. Chen, X. Tao, and Y. Sun, "Adaptive cell zooming and sleeping for green heterogeneous ultradense networks," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 2, pp. 1612–1621, feb 2018.

[11] A. El Amine, M. Iturralde, H. A. H. Hassan, and L. Nuaymi, "A distributed q-learning approach for adaptive sleep modes in 5g networks," in *2019 IEEE Wireless Communications and Networking Conference (WCNC)*, 2019.

[12] F. E. Salem, Z. Altman, A. Gati, T. Chahed, and E. Altman, "Reinforcement learning approach for advanced sleep modes management in 5g networks," in *2018 IEEE 88th Vehicular Technology Conference (VTC-Fall)*. IEEE, 2018, pp. 1–5.

[13] M. Ozturk, M. Akram, S. Hussain, and M. A. Imran, "Novel qos-aware proactive spectrum access techniques for cognitive radio using machine learning," *IEEE Access*, vol. 7, pp. 70 811–70 827, 2019.

[14] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[15] M. Ozturk, M. Jaber, and M. A. Imran, "Energy-aware smart connectivity for iot networks: Enabling smart ports," *Wireless Communications and Mobile Computing*, vol. 2018, 2018.

[16] C. J. Watkins and P. Dayan, "Q-learning," *Machine learning*, vol. 8, no. 3-4, pp. 279–292, 1992.